Probabilidade & Estatística

Sumário

Variáveis Aleatórias, Experimentos, Dados e Amostragem

- Informalmente, uma **variável aleatória** é um característico numérico de um experimento.
- Exemplo: se o experimento consiste em lançar uma moeda e anotar a face superior resultante do lançamento, podemos considerar a variável aleatória que atribui o valor numérico 1 no caso de o resultado do lançamento ser ‘cara’, e 0 se for ‘coroa’.
- Exemplo: se o experimento consiste em um estudo médico em que diversas características de um grupo de pacientes são medidas, então uma possível variável aleatória a ser considerada é, digamos, o *índice glicêmico do paciente* j, onde j indexa os pacientes que integram o estudo.

- A noção de experimento aqui é vaga, e pode contemplar, por exemplo:
  - experimentos laboratoriais: medições de quantidades físicas em um ambiente controlado.
  - dados observacionais: medições em ambientes não controlados (exemplo: dados meteorológicos).
  - pesquisas (exemplo: questionários, dados médicos).
  - séries temporais (exemplo: evolução temporal da taxa de câmbio).

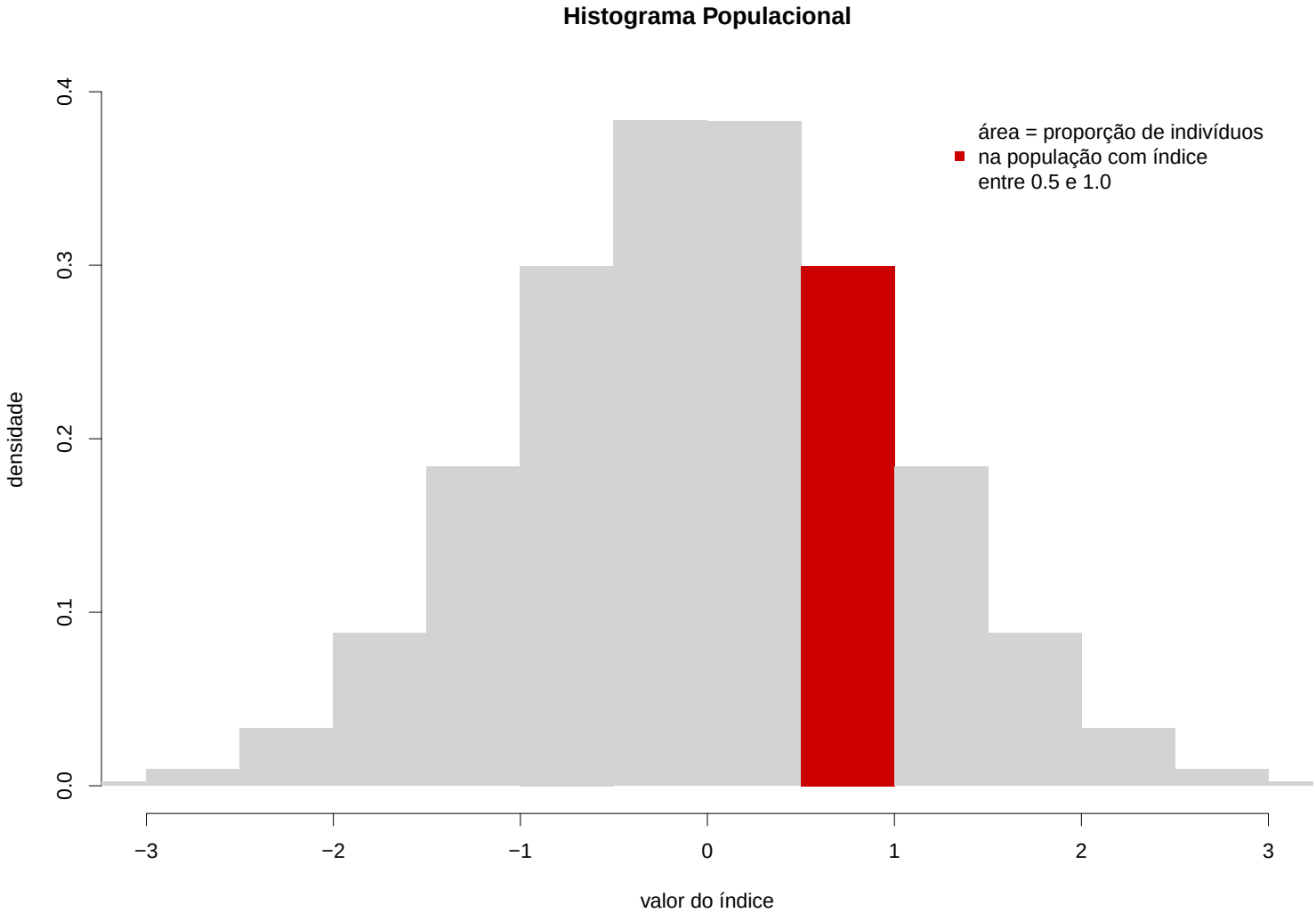
- Usualmente, denota-se abstratamente o resultado do experimento em questão pela letra grega ω. O conjunto de todos os possíveis resultados de um experimento é denotado pela letra grega Ω, chamado o *espaço amostral*.
- Por exemplo, se considerarmos o experimento anterior consistindo em lançar uma moeda e anotar a face superior resultante do lançamento, então uma possível formulação é considerar o espaço amostral Ω como sendo o conjunto constituído pelas palavras ‘cara’ e ‘coroa’.
- Nesse espaço amostral, ω pode tão somente ser igual a ‘cara’ ou então igual a ‘coroa’.

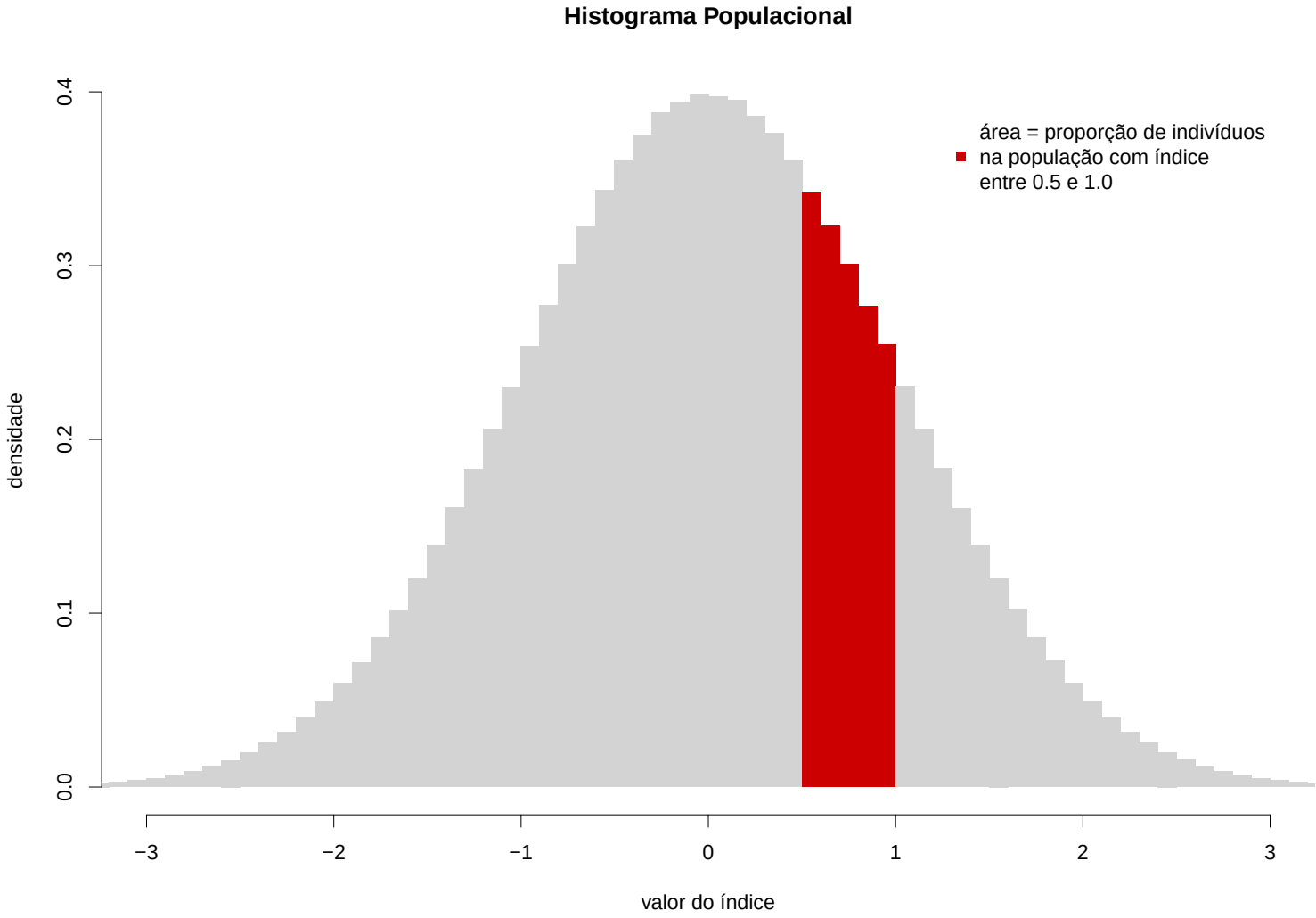
- As variáveis aleatórias associadas a um certo experimento são denotadas usualmente por letras maiúsculas, em alguns casos com um índice subscrito: X, Y, Z, X_1, X_2, $\ldots$, X_n, etc.
- É importante distinguir: *(i)* a variável aleatória, digamos X, entendida abstratamente como a 'regra' que atribui um valor numérico a cada resultado do experimento, e *(ii)* o preciso valor numérico $X(\omega)$ correspondente ao cenário em que o resultado do experimento é ω.
- No exemplo da moeda, $X(\text{'cara'}) = 1$, enquanto $X(\text{'coroa'}) = 0$.
- Formalmente, X é uma função real com domínio Ω, enquanto $X(\omega)$ é o valor numérico específico que essa função atribui ao resultado ω.

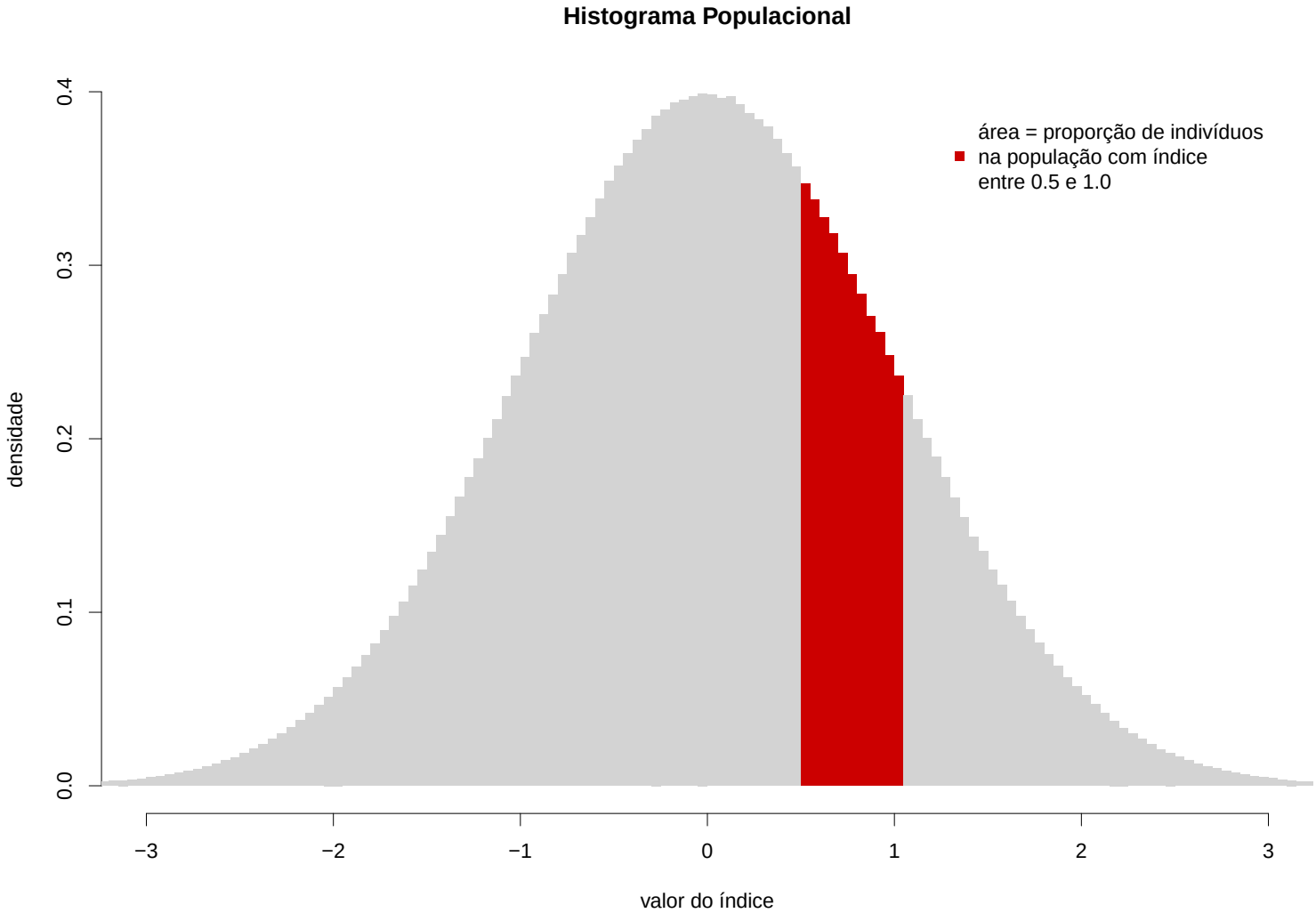
- Relacionado à noção de experimento aleatório está o conceito de amostra.
- Intuitivamente, uma amostra é uma parte de uma população.
- Por exemplo, se considerarmos como população o conjunto de todos os habitantes da região metropolitana de Porto Alegre, então uma possível amostra consiste em *selecionar*, digamos, um grupo de 100 pessoas que residem nessa região.
- No exemplo acima, o experimento consiste em selecionar (ao acaso?) $n = 100$ pessoas de uma população. Uma possível especificação é definir o espaço amostral como sendo o conjunto de todos os possíveis sorteios de 100 indivíduos dessa população.
- Exemplos de variáveis aleatórias nesse contexto: *renda mensal do j-ésimo indivíduo sorteado*, *altura do j-ésimo indivíduo sorteado*, etc.

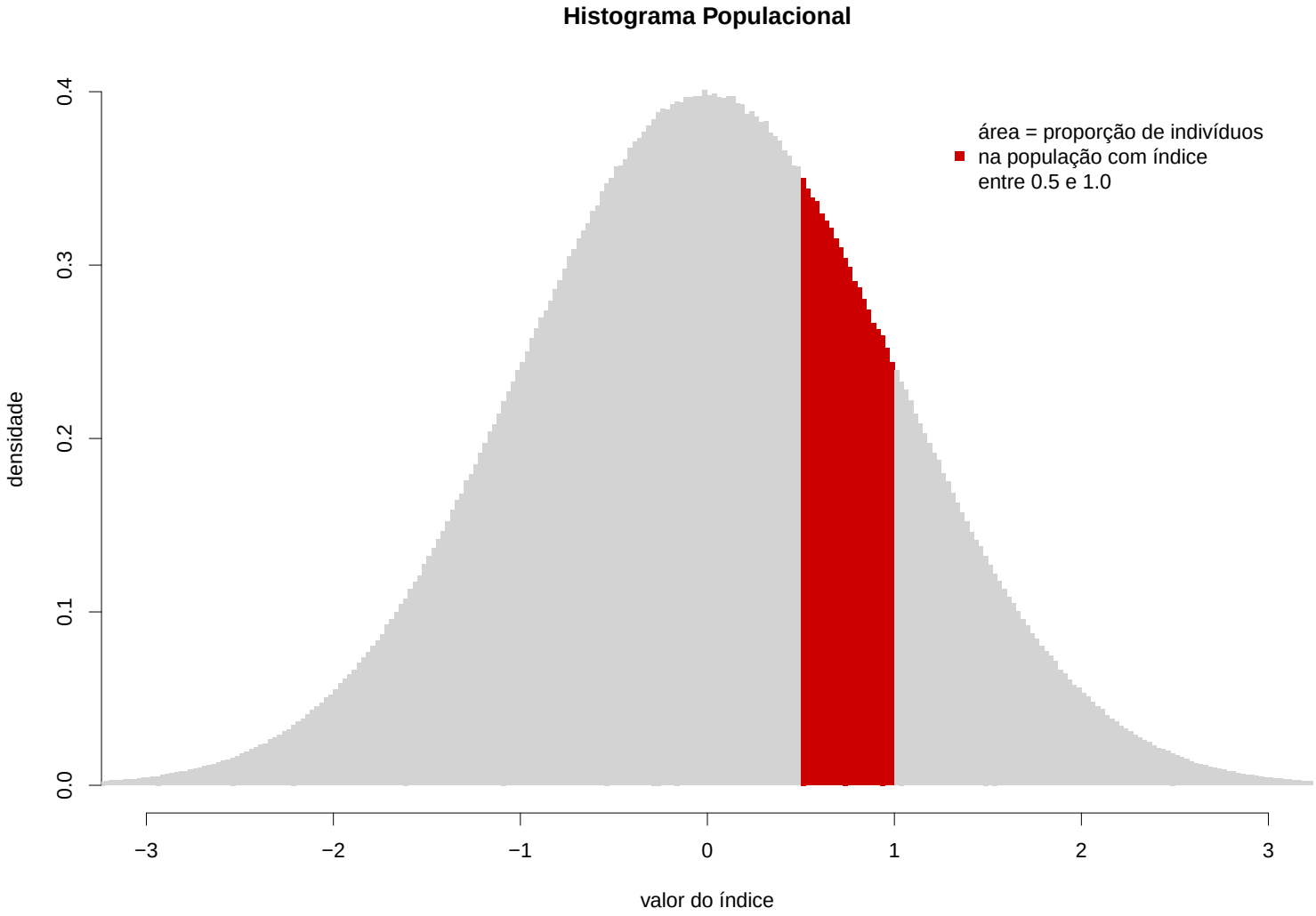
- É importante ressaltar que a noção intuitiva de amostra nem sempre é adequada. No exemplo acima, a população é finita. Será conveniente introduzirmos uma noção para uma amostra de uma população *infinita*.
- Uma maneira de apreender essa ideia é considerar uma população finita mas ‘muito grande’, e observar os histogramas dessa população.
- Nos gráficos a seguir temos os histogramas referentes a uma população de tamanho 10000000. A característica de interesse é um número que chamaremos genericamente de ‘índice’.

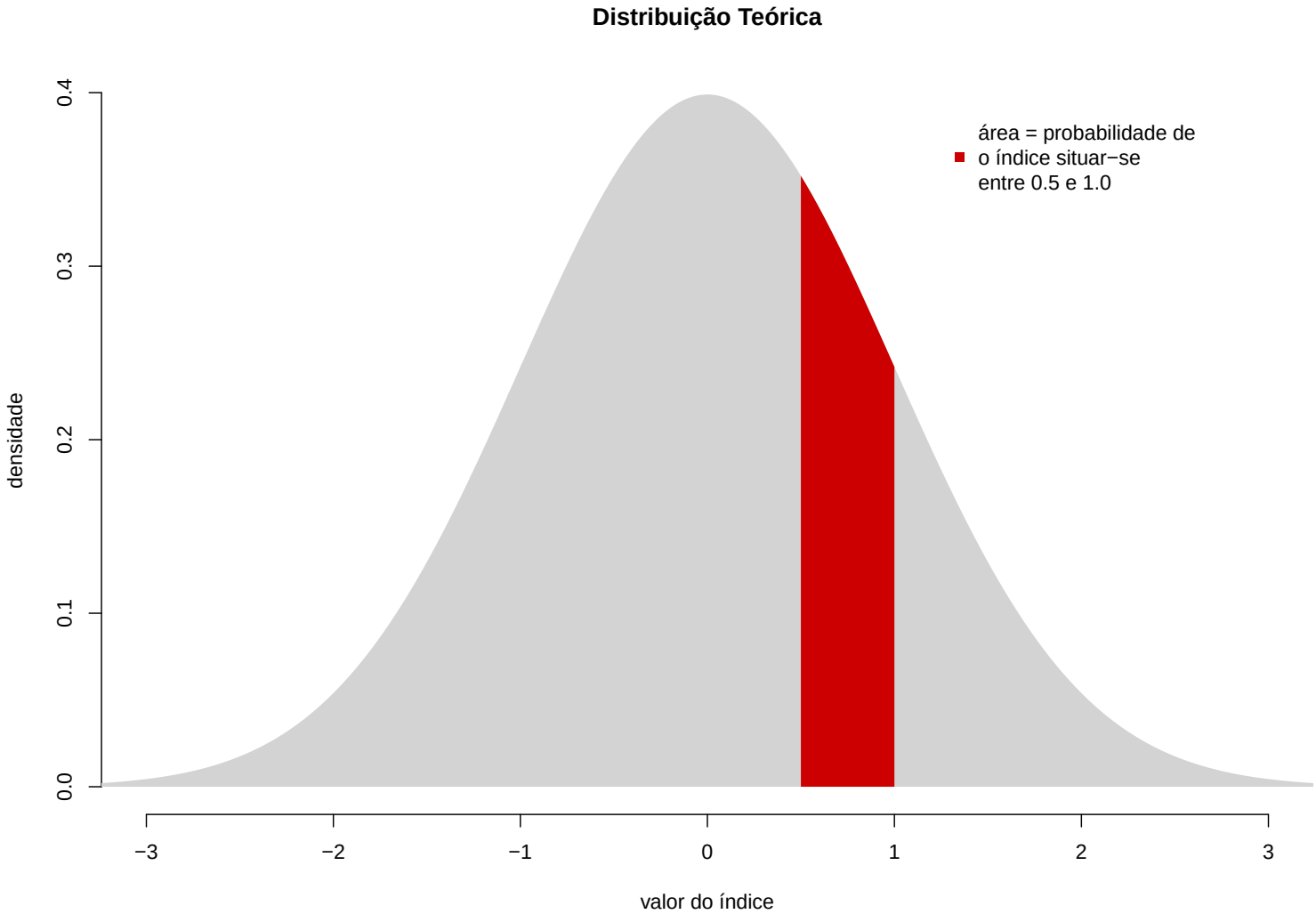
- É importante ressaltar que a noção intuitiva de amostra nem sempre é adequada. No exemplo acima, a população é finita. Será conveniente introduzirmos uma noção para uma amostra de uma população *infinita*.
- Uma maneira de apreender essa ideia é considerar uma população finita mas ‘muito grande’, e observar os histogramas dessa população.
- Nos gráficos a seguir temos os histogramas referentes a uma população de tamanho 10000000. A característica de interesse é um número que chamaremos genericamente de índice.

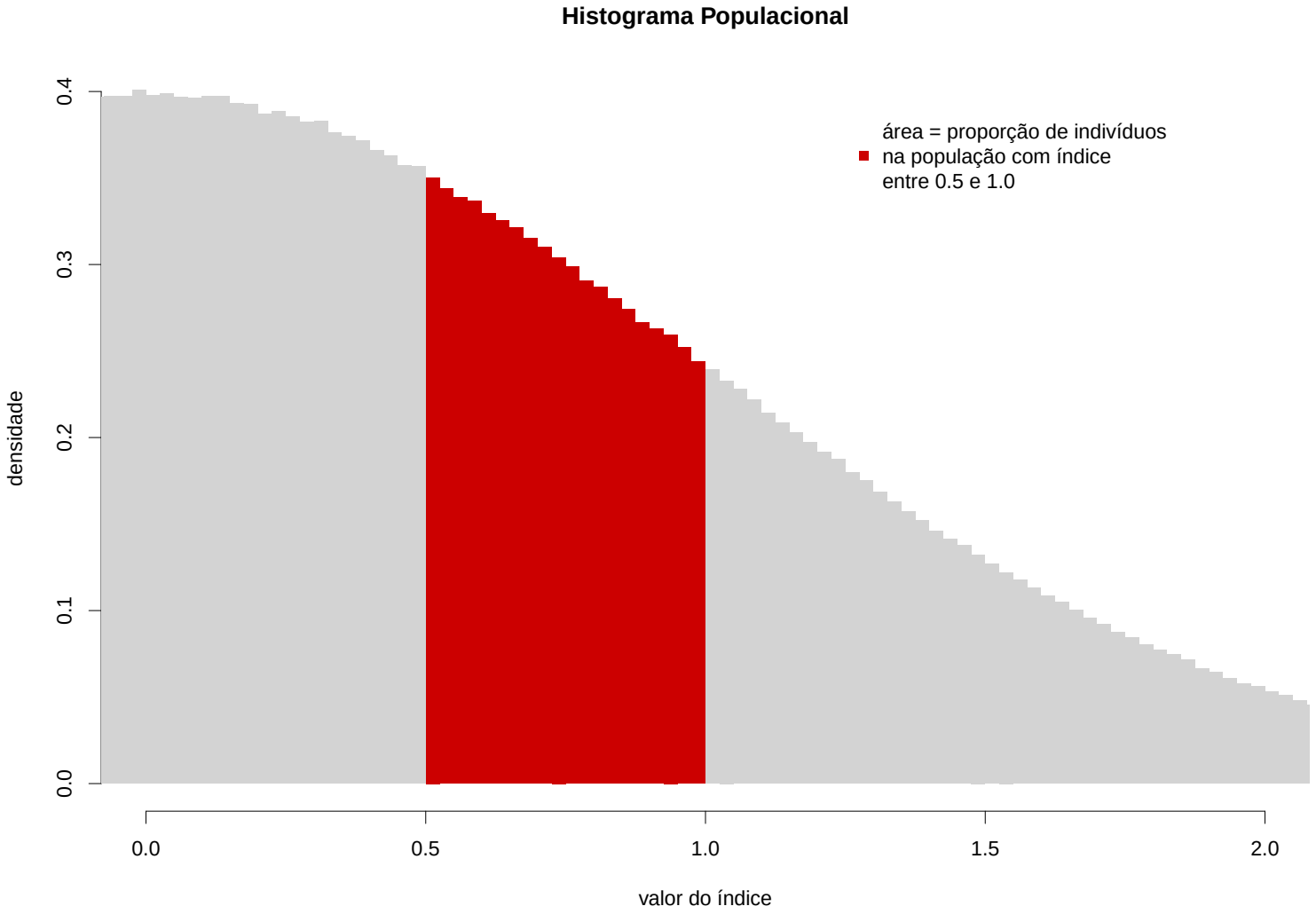

Histograma Populacional
área = proporção de indivíduos
na população com índice
entre 0.5 e 1.0
densidade
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
valor do índice

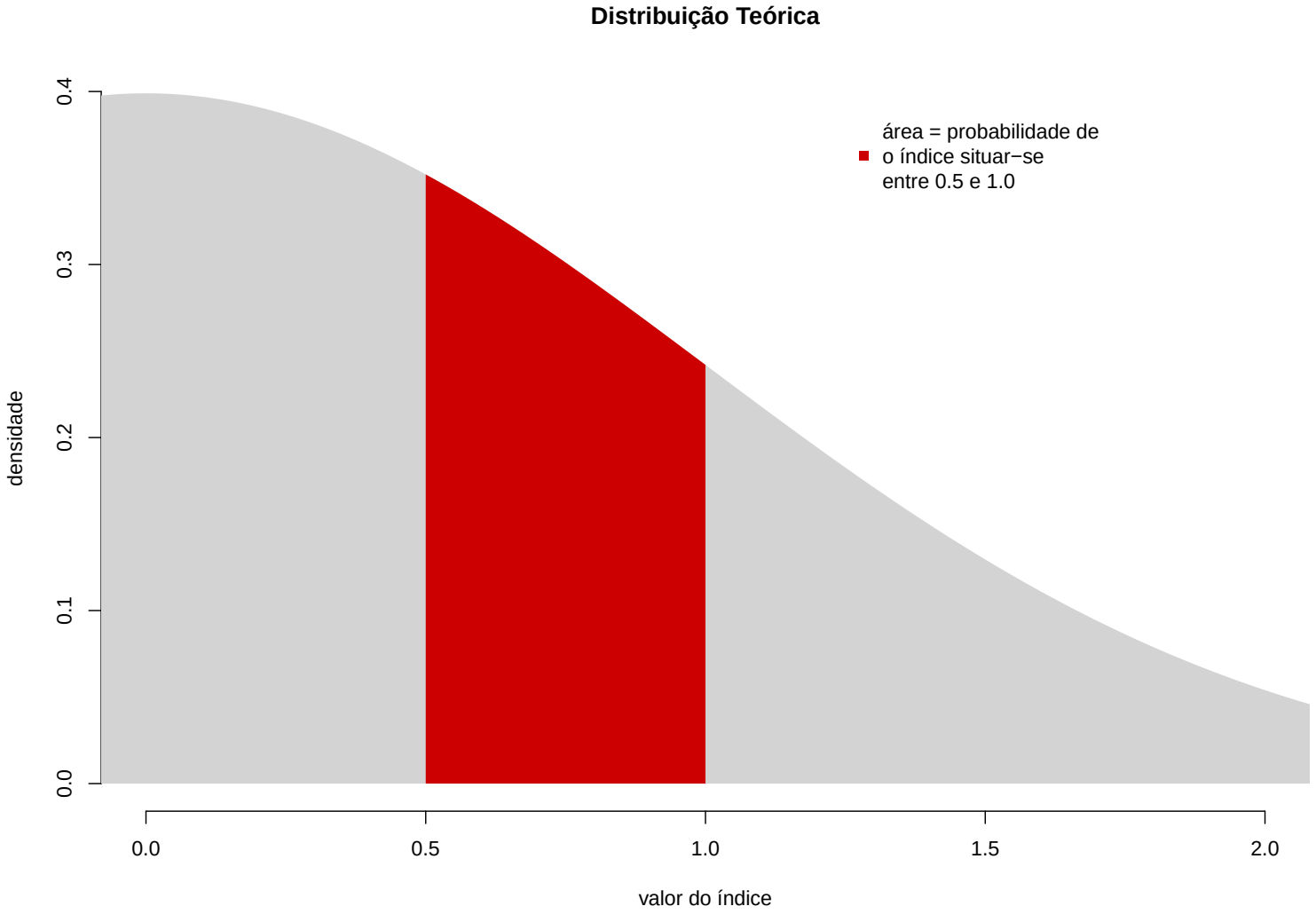

Histograma Populacional
área = proporção de indivíduos
na população com índice
entre 0.5 e 1.0
densidade
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
valor do índice


Histograma Populacional
área = proporção de indivíduos
na população com índice
entre 0.5 e 1.0
densidade
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
valor do índice


Histograma Populacional
área = proporção de indivíduos
na população com índice
entre 0.5 e 1.0
densidade
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
valor do índice


Distribuição Teórica
área = probabilidade de
o índice situar−se
entre 0.5 e 1.0
densidade
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
valor do índice


Histograma Populacional
área = proporção de indivíduos
na população com índice
entre 0.5 e 1.0
densidade
0.0
0.1
0.2
0.3
0.4
0.0
0.5
1.0
1.5
2.0
valor do índice


Distribuição Teórica
área = probabilidade de
o índice situar−se
entre 0.5 e 1.0
densidade
0.0
0.1
0.2
0.3
0.4
0.0
0.5
1.0
1.5
2.0
valor do índice

- Nos histogramas acima, a área de cada uma das barras indica a *proporção* de indivíduos da população que possuem o valor do índice entre o extremo inferior e o extremo superior que delimitam a base da barra em questão.
- A densidade, em contrapartida, deve ser entendida como uma *distribuição teórica*. A curva que delimita a área em cinza no gráfico anterior é dada pela fórmula matemática

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

onde x é qualquer número real.

- Por ser um modelo teórico, a densidade não depende da população. Ao contrário: o valor do índice é que deve ser entendido como uma quantidade aleatória cuja 'lei' de probabilidade é dada pela fórmula acima, no seguinte sentido: se a e b são dois números quaisquer com $a < b$, então a probabilidade de o índice situar-se entre a e b é igual a

$$\int_a^b \varphi(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx$$

isto é, à área compreendida sob a curva da densidade e entre a e b.

- No exemplo acima, a área em vermelho indica a probabilidade de o índice situar-se entre 0.5 e 1.0.
- Esse valor pode ser interpretado como sendo a probabilidade (teórica) de que o índice de um indivíduo sorteado ao caso da população tenha seu valor compreendido entre 0.5 e 1.0.
- Convém memorizar:
  - eixo horizontal: possíveis valores de uma variável aleatória (no exemplo, o 'índice').
  - áreas = probabilidades.
- É importante perceber que no exemplo acima o modelo teórico é uma boa *aproximação* para a distribuição populacional. Isto é, assumindo que os valores do índice de todos os indivíduos de uma população finita sejam conhecidos, então em sucessivos refinamentos do histograma as probabilidades de ocorrência de certos eventos se aproximam das probabilidades calculadas a partir da fórmula dada pelo modelo teórico.

- Evidentemente, o exemplo acima é apenas uma simulação computacional, onde de fato conhecemos todas as informações de uma população finita (no caso, o valor numérico do ‘índice’ de cada um dos 10000000 de indivíduos). Também sabemos a priori qual é o modelo teórico correto!
- É importante ressaltar que em geral isso não ocorre. O que se tem é um conjunto de dados (uma amostra) a partir da qual se pretende inferir propriedades da população.
- Todavia, muitas vezes é conveniente passar a interpretar o termo ‘população’ como sendo algum modelo teórico desconhecido — mesmo quando se sabe que a amostra que temos em mãos foi coletada a partir de uma população finita.
- O objetivo da inferência estatística passa a ser o de descobrir (ou propor) um modelo teórico adequado para aquele conjunto de dados.

- Para fixar ideias: **população** é um conjunto de *unidades* sobre as quais se quer obter informação (*unidades* é um termo usado vagamente para designar objetos, itens, pessoas, animais).
- Quando esse conjunto de unidades é finito, a intuição ‘funciona’. Quando for infinito (ver o exemplo a seguir), a interpretação se torna um tanto mais vaga.
- **Amostra** é uma parte observável/observada da população.

- Um exemplo clássico em que a ‘população’ não é constituída por um número finito de indivíduos é o seguinte: digamos que uma certa tecnologia seja utilizada para produzir lâmpadas.
- Antes de as lâmpadas serem comercializadas, essa tecnologia será testada da seguinte forma: uma *amostra* de 100 lâmpadas produzidas de acordo com a tecnologia em questão será produzida, e o tempo de vida de cada uma delas será medido.
- Para todos os efeitos, cada lâmpada é uma réplica perfeita das demais. Quaisquer diferenças entre os tempos de vida dessas lâmpadas se devem a fatores intrinsicamente aleatórios.

- Os tempos de vida dessas 100 lâmpadas amostradas podem ou não nos dar um indicativo de qual é o tempo de vida esperado de uma lâmpada produzida de acordo com a tecnologia em questão. Todavia, por mais que se produzam lâmpadas adicionais jamais será possível conhecer a ‘população’ da qual estamos amostrando.
- De fato, aqui parece ser mais conveniente assumir que o tempo de vida de cada uma dessas lâmpadas seja uma quantidade aleatória, governada por algum modelo teórico o qual traduz toda a estrutura de probabilidade que se manifesta nos experimentos.

- Por exemplo, se quisermos responder à pergunta “qual é a probabilidade de uma lâmpada produzida de acordo com essa tecnologia ter um tempo de vida superior a 3×10^4 segundos?”, claramente não faz sentido calcular a razão

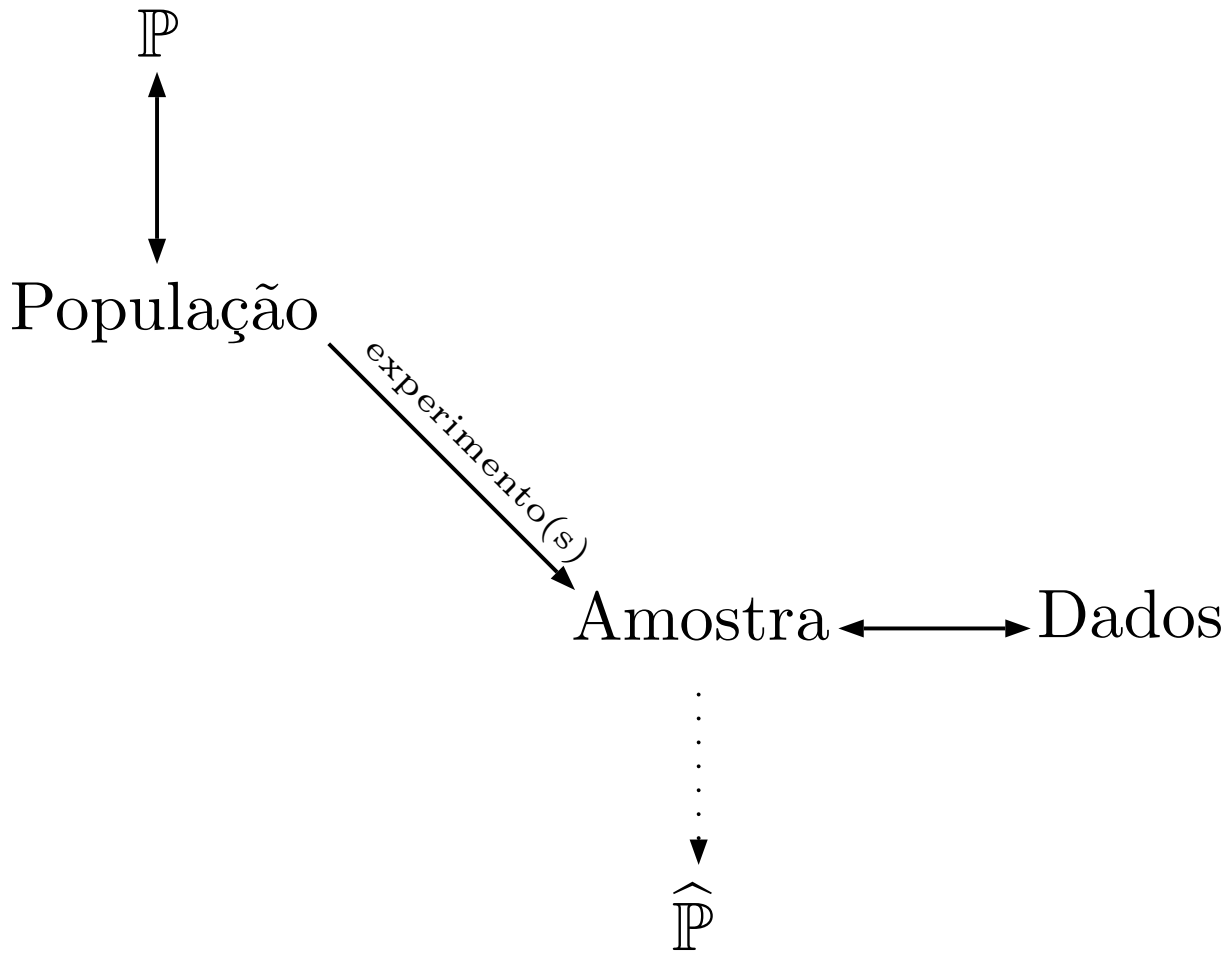
$$\frac{\text{número de lâmpadas cujo tempo de vida é maior ou igual a } 3 \times 10^4 s}{\text{número total de lâmpadas}}$$

já que o denominador sequer faz sentido (na verdade, nem o numerador faz sentido).

- Assim, devemos admitir que a resposta à questão acima seja dada em termos de uma propriedade latente, a qual será observada (ao menos aproximadamente) se repetirmos o experimento “medir o tempo de vida de uma lâmpada” um grande número de vezes.
- Por exemplo, se de acordo com algum modelo teórico a probabilidade acima for, digamos, igual a 0.5, então esperaremos que em nossa amostra inicial de 100 lâmpadas, aproximadamente 50 delas tenham durado mais do que 3×10^4 segundos.
- Nada impede, todavia, que em nossa amostra específica essa proporção tenha sido violada. De fato, podemos admitir até mesmo que todas as lâmpadas queimarem em menos de 10 segundos seja um evento possível. A ideia é que o modelo teórico dá a caracterização completa de quão improvável é tal evento.

- Menos formal e mais concretamente, o que geralmente temos em mãos é algum conjunto de dados, dispostos por exemplo em uma planilha. Esse conjunto de dados pode ser considerado, *latu sensu*, como sendo nossa amostra. Vamos assumir, por simplicidade, que sejam dados numéricos.
- Para compreender qual é o objetivo da inferência estatística, é conveniente interpretar esse conjunto de dados (isto é, essa amostra) tendo em vista as seguintes hipóteses:
  - Esse conjunto de dados poderia ser diferente.
  - Se fosse diferente, essa variabilidade é proveniente de algum 'mecanismo aleatório'.
- Nos termos anteriores 'ser diferente' significa que o resultado do 'experimento', ω, é outro. O mecanismo aleatório é o modelo teórico através do qual ω é 'sorteado'.
- O nosso objetivo passa a ser o de inferir qual é o mecanismo que está 'gerando' os dados.

$\mathbb{P}$ = estrutura probabilística da população



Formas de seleção de uma amostra

- A maneira pela qual o pesquisador obtém os dados (i.e, a amostra) é extremamente importante. As mais conhecidas são *amostragem* e *planejamento de experimentos*.
- Podemos dividir os procedimentos científicos de obtenção de dados amostrais em três grupos:
    - Levantamentos amostrais;
    - Planejamento de experimentos;
    - Levantamentos observacionais.

- **Levantamentos amostrais**: correspondem àquele cenário em que a amostra é obtida de uma população bem definida, por meio de procedimentos bem definidos e controlados pelo pesquisador. Pode ser subdividida em dois subgrupos:
    1. Levantamentos probabilísticos;
    2. Levantamentos não probabilísticos.
- Levantamentos probabilísticos: são aquelas técnicas que utilizam mecanismos aleatórios de seleção dos elementos de uma amostra, atribuindo a cada um deles uma probabilidade, conhecida a priori, de pertencer à amostra.
- Levantamentos não probabilísticos: procedimentos nos quais a amostra é selecionada intencionalmente (por exemplo, com o auxílio de um especialista), ou em que a amostra é composta por voluntários.

- **Planejamento de experimentos**: o principal objetivo dessa forma de obtenção de dados é o de analisar o efeito de uma variável sobre outra. Requer interferências do pesquisador sobre o ambiente em estudo (população), bem como o controle de fatores externos, com o intuito de medir o efeito desejado.
- **Levantamentos observacionais**: nesse caso, os dados são coletados sem que o pesquisador tenha controle sobre as informações obtidas. Séries temporais são um exemplo típico.
- Obs: quando o objeto de estudo é uma série temporal, o modelo subjacente é o de processo estocástico. A trajetória efetivamente observada é entendida como apenas uma dentre todas as trajetórias possíveis.

- Dentro do esquema de levantamentos amostrais probabilísticos, destaca-se a *amostragem aleatória simples*.
- No caso de uma população finita, esse procedimento é análogo a escrever cada elemento da população em um cartão, misturar os cartões em uma urna e sortear tantos cartões quanto desejarmos na amostra.
- Definimos uma *amostra aleatória simples* de tamanho n de uma população como sendo um conjunto $X_1, X_2, \ldots, X_n$ de variáveis aleatórias independentes e identicamente distribuídas (iid).
- Isso quer dizer que a estrutura de probabilidade de cada X_i é a mesma, e que eventos associados a qualquer um dos X_i não influenciam a probabilidade de ocorrência de eventos associados às demais variáveis da amostra.

$$\mathbb{P}(X_i \leq t) = \mathbb{P}(X \leq t), \qquad \text{(não depende de } i\text{)}$$

$$\mathbb{P}(X_1 \leq t_1, X_2 \leq t_2, \ldots X_n \leq t_n) = \mathbb{P}(X_1 \leq t_1)\mathbb{P}(X_2 \leq t_2) \cdots \mathbb{P}(X_n \leq t_n)$$

Parâmetros, Estatísticas e Estimadores

Parâmetros

- Anteriormente falamos em uma “estrutura probabilística da população”. Muitas vezes convém caracterizar essa expressão em termos de *parâmetros*.
- **Parâmetros** são valores numéricos fixos associadas a uma população (ou a um modelo teórico).
- Um exemplo clássico de parâmetro é a *média populacional* de alguma característica numérica atribuível a cada unidade de uma população. Por motivos mnemônicos, guarde a expressão ‘valor esperado populacional’. Falar em média amostral vs média populacional pode confundir.
- Se considerarmos a característica numérica acima como sendo alguma variável aleatória X, então a média populacional é às vezes chamada de esperança matemática de X (ou simplesmente esperança de X, ou valor esperado de X, etc.)

- Exemplo: se considerarmos como população o conjunto de todos os habitantes da região metropolitana de Porto Alegre, e se a característica numérica de interesse for, digamos, a altura, então podemos a princípio calcular a altura média populacional dos habitantes dessa região (para isso, teríamos que medir a altura de todos esses habitantes)

- Por outro lado, se considerarmos nosso exemplo anterior de uma tecnologia de produção de lâmpadas, então como dar sentido a uma expressão do tipo “tempo médio de vida de uma lâmpada produzida de acordo com essa tecnologia”?
- Nesse contexto é conveniente considerar o tempo de vida de uma lâmpada (produzida de acordo com a tecnologia em questão) como sendo uma variável aleatória X, cuja lei de probabilidade é governada por algum modelo teórico. O valor esperado de X, denotado por

$$\mathbb{E}X$$

pode ser entendido, aqui, como uma propriedade latente (um parâmetro de locação) que se manifesta, aproximadamente, em repetições do experimento “medir o tempo de vida de uma lâmpada”.

- Como calcular o valor esperado (teórico) de X?
- Se X for variável aleatória contínua com função densidade de probabilidade f_X, usamos a fórmula

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

- Se X for variável aleatória discreta com função massa de probabilidade p_X, usamos a fórmula

$$\mathbb{E}X = \sum_i x_i\, p_X(x_i),$$

onde $x_1, x_2, \ldots$ são os possíveis valores assumidos por X.

- Essas fórmulas dependem da estrutura de probabilidade *populacional* (a densidade ou a massa de probabilidade). Em um trabalho aplicado, isso é *desconhecido*.
- Isso quer dizer que *não é possível* obter os valores de parâmetros a partir de dados de uma amostra!
- Não confunda aquilo que você pode calcular a partir dos dados, com aquilo que você só pode calcular se conhecer toda a estrutura probabilística da população.

Estatísticas e Estimadores

- Uma **estatística** é uma variável aleatória construída a partir de uma amostra. Em outras palavras, uma estatística é qualquer função dos dados: se $X_1, X_2, \ldots, X_n$ é uma amostra iid, então qualquer função da forma

$$T(\omega) = g(X_1(\omega), \ldots, X_n(\omega)),$$

  onde g é uma função qualquer de n variáveis, é uma estatística.
- Estatísticas também são chamadas **estimadores**.
- Observe que, por definição, estatísticas (e estimadores) são antes de tudo variáveis aleatórias!

- Dois exemplos clássicos de estatísticas/estimadores são a média amostral,

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  e a variância amostral

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

- Sendo mais preciso, se o resultado do experimento é ω, então

$$\bar{X}_n(\omega) = \frac{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i(\omega)$$

  e

$$S_n^2(\omega) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i(\omega) - \bar{X}_n(\omega))^2.$$

- Como um estimador é uma variável aleatória, estamos autorizados a nos indagar sobre a estrutura probabilística a ele associada.
- Por exemplo, poderíamos estar interessados em calcular

$$\mathbb{P}(\bar{X}_n > 2) = ? \qquad \mathbb{E}(\bar{X}_n) = ? \qquad \mathrm{Var}(\bar{X}_n) = ?$$

Distribuição Amostral de um Estimador

- A intuição nos diz que se T é um estimador da forma

$$T(\omega) = g(X_1(\omega), \ldots, X_n(\omega)),$$

  então a estrutura probabilística associada à população da qual obtivemos a amostra $X_1, \ldots, X_n$ deveria influenciar em algum sentido a estrutura probabilística do estimador em questão.
- De fato,

$$\mathbb{P}(T > 2) = \mathbb{P}(g(X_1, \ldots, X_n) > 2).$$

- Veremos que, nesse caso, a intuição está correta.

- Lembrando: uma amostra aleatória de tamanho n de uma variável aleatória X é uma coleção $X_1, X_2, \ldots, X_n$ de variáveis aleatórias independentes e identicamente distribuídas (iid), tendo a mesma distribuição que X. Quer dizer, se X tem função de distribuição acumulada F_X, então

$$\mathbb{P}(X_i \leq x) = \mathbb{P}(X \leq x) = F_X(x)$$

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots X_n \leq x_n) = F_X(x_1)F_X(x_2)\cdots F_X(x_n)$$

- Equivalentemente, dizemos que $X_1, X_2, \ldots, X_n$ é uma amostra aleatória da população F_X.

- A **distribuição amostral** (teórica) de um estimador é a lei de probabilidade associada a esse estimador (visto como uma variável aleatória).
- Exemplo: digamos que a população seja composta pelos números 1, 3, 5, 5, e 7. Se considerarmos amostras aleatórias (com reposição) de tamanho 2 dessa população, então as possíveis amostras que obteremos são (na lista a seguir já calculamos o valor da média amostral correspondente a cada amostra)

Possíveis amostras de tamanho $n = 2$ da população $1, 3, 5, 5, 7$, e o respectivo valor da média amostral.

$X_1(\omega) = 1$	$X_2(\omega) = 1$	$\bar{X}_2(\omega) = 1$
$X_1(\omega) = 1$	$X_2(\omega) = 3$	$\bar{X}_2(\omega) = 2$
$X_1(\omega) = 1$	$X_2(\omega) = 5$	$\bar{X}_2(\omega) = 3$
$X_1(\omega) = 1$	$X_2(\omega) = 7$	$\bar{X}_2(\omega) = 4$
$X_1(\omega) = 3$	$X_2(\omega) = 1$	$\bar{X}_2(\omega) = 2$
$X_1(\omega) = 3$	$X_2(\omega) = 3$	$\bar{X}_2(\omega) = 3$
$X_1(\omega) = 3$	$X_2(\omega) = 5$	$\bar{X}_2(\omega) = 4$
$X_1(\omega) = 3$	$X_2(\omega) = 7$	$\bar{X}_2(\omega) = 5$
$X_1(\omega) = 5$	$X_2(\omega) = 1$	$\bar{X}_2(\omega) = 3$
$X_1(\omega) = 5$	$X_2(\omega) = 3$	$\bar{X}_2(\omega) = 4$
$X_1(\omega) = 5$	$X_2(\omega) = 5$	$\bar{X}_2(\omega) = 5$
$X_1(\omega) = 5$	$X_2(\omega) = 7$	$\bar{X}_2(\omega) = 6$
$X_1(\omega) = 7$	$X_2(\omega) = 1$	$\bar{X}_2(\omega) = 4$
$X_1(\omega) = 7$	$X_2(\omega) = 3$	$\bar{X}_2(\omega) = 5$
$X_1(\omega) = 7$	$X_2(\omega) = 5$	$\bar{X}_2(\omega) = 6$
$X_1(\omega) = 7$	$X_2(\omega) = 7$	$\bar{X}_2(\omega) = 7$

- População de tamanho 5, amostra de tamanho 2, e as coisas já ficaram complicadas. . .
- O que importa é que aqui a média amostral é uma variável aleatória discreta, cujos possíveis valores são $1, 2, \ldots, 7$.
- Como obter a função massa de probabilidade de média amostral?
- Quer dizer, como obter $\mathbb{P}(\bar{X}_2 = i)$, $i = 1, 2 \ldots, 7$?
- Conforme havíamos intuído, isso vai depender da "estrutura probabilística" das variáveis aleatórias X_1 e X_2.

- Sabemos que

$$\begin{aligned}
\mathbb{P}(X_1 = 1) &= \mathbb{P}(X_2 = 1) = 1/5 \\
\mathbb{P}(X_1 = 3) &= \mathbb{P}(X_2 = 3) = 1/5 \\
\mathbb{P}(X_1 = 5) &= \mathbb{P}(X_2 = 5) = 2/5 \\
\mathbb{P}(X_1 = 7) &= \mathbb{P}(X_2 = 7) = 1/5.
\end{aligned}$$

- Além disso,

$$\mathbb{P}(X_1 = i,\, X_2 = j) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = j).$$

- Agora já temos todas as informações necessárias para calcular o valor esperado da média amostral:

$$\mathbb{P}(\bar{X}_2 = 1) = 1/25$$
$$\mathbb{P}(\bar{X}_2 = 2) = 2/25$$
$$\mathbb{P}(\bar{X}_2 = 3) = 5/25$$
$$\mathbb{P}(\bar{X}_2 = 4) = 6/25$$
$$\mathbb{P}(\bar{X}_2 = 5) = 6/25$$
$$\mathbb{P}(\bar{X}_2 = 6) = 4/25$$
$$\mathbb{P}(\bar{X}_2 = 7) = 1/25$$

- Disso segue que

$$\mathbb{E}(\bar{X}_2) = \sum_{x=1}^{7} x\,\mathbb{P}(\bar{X}_2 = x) = 4.2$$

- Curiosamente,

$$\frac{1+3+5+5+7}{5} = 4.2$$

  e também

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) = 4.2$$

- Podemos calcular, analogamente,

$$\mathrm{Var}(\bar{X}_2) = \sum_{x=1}^{7} (x - \mathbb{E}(\bar{X}_2))^2 \times \mathbb{P}(\bar{X}_2 = x) = \cdots = 2.08$$

- Curiosamente,

$$\frac{(1-4.2)^2+(3-4.2)^2+(5-4.2)^2+(5-4.2)^2+(7-4.2)^2}{5} = 4.16 = 2 \times 2.08$$

  e também

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 4.16 = 2 \times 2.08$$

- . . . surge um padrão?

- O exemplo acima é útil mais como forma de acostumar o leitor à ideia de que a média amostral é uma variável aleatória, e de que a expressão “média da média” faz sentido (embora, por questões estéticas/mnemônicas, seja mais apropriado falar em “valor esperado da média amostral”, etc)
- Entretanto, o exemplo acima é extremamente simples, já que a população consistia de apenas 5 indivíduos (e a variável aleatória em questão somente podia assumir 4 valores), e consideramos somente amostras aleatórias de tamanho 2.
- Apesar da simplicidade do exemplo, vimos que obter a distribuição de probabilidade desse estimador pelo “método direto” já tornou as coisas um tanto complicadas.

Distribuição amostral do estimador “média amostral”

- O seguinte resultado mostra, ao menos, que os dois “fatos curiosos” que apareceram no exemplo não são exceções mas sim a regra.
- **Teorema**: se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória da variável aleatória X, então

$$\mathbb{E}(\bar{X}_n) = \mathbb{E}(X), \qquad \mathrm{Var}(\bar{X}_n) = \frac{\mathrm{Var}(X)}{n}.$$

  Isto é: o valor esperado do estimador ‘média amostral’ é igual ao valor esperado da variável aleatória da qual a amostra foi tomada
- Isto é: o valor esperado do estimador ‘média amostral’ é igual ao valor esperado da variável aleatória da qual a amostra foi tomada.
- A variância do estimador ‘média amostral’ é igual à variância (populacional) da variável aleatória da qual a amostra foi tomada, dividida pelo tamanho da amostra.

- O valor esperado e a variância são parâmetros, respectivamente, de locação (posição) e de dispersão.
- Portanto, o Teorema acima nos dá duas informações sobre a estrutura probabilística associada ao estimador ‘média amostral’.
- É possível refinar o resultado acima. Por exemplo, poderíamos estar interessados em calcular

$$\mathbb{P}(\bar{X}_2 \leq 2).$$

- O Teorema seguinte (um dos resultados mais importantes da Teoria da Probabilidade), nos diz que, não importa qual seja a população da qual se toma a amostra, se essa amostra for suficientemente grande então a variável aleatória “média amostral” segue uma distribuição aproximadamente normal.

- **Teorema Central do Limite**: se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória da variável aleatória X, e o tamanho da amostra n for suficientemente grande, então

$$\mathbb{P}(\bar{X}_n \leq x) \cong \mathbb{P}(Z \leq x),$$

onde Z é uma variável aleatória Normal com $\mathbb{E}Z = \mathbb{E}X$ e $\mathrm{Var}(Z) = \mathrm{Var}(X)/n$.

- A importância da propriedade acima reside no fato de que ela vale com total generalidade: as peculiaridades da distribuição populacional (seja ela finita ou um modelo teórico) têm pouca influência no comportamento probabilístico da variável aleatória "média amostral". Para amostras grandes, essa variável aleatória é sempre aproximadamente Normal.
- **Corolário**: Seja X uma variável aleatória qualquer com $\mathbb{E}X = \mu$ e $\mathrm{Var}(X) = \sigma^2$. Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de X, e n for suficientemente grande, então a variável aleatória

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

tem distribuição Normal com valor esperado igual a zero e variância igual a 1.

Distribuição amostral de uma proporção

- Digamos que estamos interessados em estimar, a partir de uma amostra, a probabilidade de ocorrência de um certo evento.
- No caso de uma população finita, essa probabilidade é dada pela proporção de indivíduos na população que possuem certa característica de interesse, em relação ao total de indivíduos na população. Por exemplo, a proporção de indivíduos que votará em determinado candidato numa eleição.
- Caso a população seja entendida como um modelo teórico, então a probabilidade de ocorrência de um evento é uma propriedade latente, que manifesta-se quando repetimos um "experimento" inúmeras vezes.

- Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma população (finita ou teórica), estimamos naturalmente a probabilidade teórica (proporção populacional) através da proporção de elementos na amostra que satisfizerem à característica de interesse.
- Exemplo: digamos que estamos interessados em saber qual é a probabilidade de uma lâmpada, produzida de acordo com certa tecnologia, ter um tempo de vida maior do que 3×10^4 segundos. Uma maneira de estimar essa probabilidade é tomando uma amostra de n lâmpadas, executar o experimento 'medir o tempo de vida de cada uma das n lâmpadas', e contar quantas delas tiveram um tempo de vida, nesse particular experimento, superior a 3×10^4 segundos.

- Vejamos que o estimador ‘proporção amostral’ é um caso especial do estimador ‘média amostral’.
- Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma população (finita ou teórica), defina as variáveis aleatórias Z_i, para i indo de 1 até n, da seguinte forma: $Z_i(\omega) = 1$ se $X_i(\omega)$ tem a característica de interesse, e $Z_i(\omega) = 0$ caso contrário.
- Claramente, as variáveis aleatórias Z_i têm distribuição Bernoulli com parâmetro p, onde p é a probabilidade / proporção populacional da característica de interesse:

$$p = \mathbb{P}(X \text{ possuir a característica de interesse}).$$

- Sendo $\widehat{p}$ o estimador 'proporção de elementos na amostra possuindo a característica de interesse', fica claro que $\widehat{p} = \bar{Z}_n$, isto é

$$\widehat{p}(\omega) = \frac{Z_1(\omega) + \cdots + Z_n(\omega)}{n}.$$

- Quer dizer, proporções amostrais são apenas um tipo especial de média amostral.
- Lembre que se Z_i tem distribuição Bernoulli com parâmetro p, então

$$\mathbb{E}(Z_i) = p, \qquad \mathrm{Var}(Z_i) = p(1-p).$$

- Portanto, pelos resultados discutidos anteriormente para estimadores que são médias amostrais, concluímos que

$$\mathbb{E}(\widehat{p}) = p, \qquad \mathrm{Var}(\widehat{p}) = \frac{p(1-p)}{n}.$$

  e que, para amostras grandes, o estimador 'proporção amostral' é uma variável aleatória com distribuição aproximadamente Normal.

Propriedades de estimadores

- Lembrando: um **estimador** é uma variável aleatória, digamos T, cujo valor depende do resultado ω de um experimento somente através dos valores $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ de uma amostra, isto é

$$T(\omega) = g(X_1(\omega), X_2(\omega), \ldots, X_n(\omega)).$$

- Costuma-se fazer a distinção entre o **estimador** T (entendido como variável aleatória) e **estimativa**, que é o valor numérico $T(\omega)$ assumido por T em uma particular amostra, quando o resultado do experimento é ω.

- Seja θ um parâmetro (isto é, algum valor numérico fixo associado à população).
- Dizemos que um estimador T é **não-viesado** para o parâmetro θ se

$$\mathbb{E}(T) = \theta.$$

- Quer dizer, um estimador é não-viesado para θ se seu valor esperado (teórico) coincide com θ. Um pouco mais informalmente: um estimador é não-viesado para θ se, em média, acerta o valor de θ.
- Se um estimador for viesado, chamamos de **viés** à quantidade

$$V(T; \theta) = \mathbb{E}(T) - \theta.$$

- A média amostral é um exemplo de estimador não-viesado (para o parâmetro 'média populacional'.

- Dizemos que uma sequência T_n de estimadores é **fortemente consistente** para o parâmetro θ se

$$\mathbb{P}(T_n \to \theta \text{ ao } n \to \infty) = 1$$

- O principal exemplo de estimador consistente é a média amostral (um resultado teórico importante, chamado Lei dos Grandes Números, garante isso).
- Intuitivamente, consistência significa que a sequência de estimadores se aproxima do parâmetro sendo estimado à medida que o tamanho da amostra aumenta, e que esse é um evento certo.

- Se T e T' são dois estimadores não-viesados de um mesmo parâmetro θ, então dizemos que T é **mais eficiente** do que T' se

$$\mathrm{Var}(T) < \mathrm{Var}(T').$$

- Assim, um estimador é mais eficiente do que outro se apresenta menor variabilidade / dispersão.
- Atenção: só vamos falar em eficiência no caso em que estivermos comparando *dois estimadores não-viesados de um mesmo parâmetro*.

- Se T é um estimador para o parâmetro θ, definimos o erro quadrático médio (do estimador T em relação a θ) pela quantidade

$$\mathrm{EQM}(T;\,\theta) = \mathbb{E}\Big\{(T-\theta)^2\Big\}.$$

- Podemos, alternativamente, escrever

$$\mathrm{EQM}(T;\,\theta) = \mathrm{Var}(T) + V(T;\,\theta).$$

Intervalos de confiança

- Até aqui, consideramos maneiras de fazer inferências pontuais sobre parâmetros.
- Por exemplo: digamos que o resultado de um experimento tenha sido ω, de forma que tenhamos uma amostra $X_1(\omega), \ldots, X_n(\omega)$ da variável aleatória X. Podemos propor uma estimativa pontual $\bar{X}_n(\omega)$ para a média populacional $\mathbb{E}(X)$.
- Nossa estimativa, por mais que esteja próxima do parâmetro sendo estimado, quase certamente será diferente do valor numérico exato desse parâmetro.
- Todavia, podemos refinar nosso procedimento de inferência, utilizando o Teorema Central do Limite, para propor certos intervalos que, com grande probabilidade, contenham o parâmetro sendo estimado.

- A ideia é a seguinte: o Teorema Central do Limite nos garante que, se $X_1, X_2, \ldots, X_n$ é uma amostra iid da variável aleatória X, então

$$\bar{X}_n \sim N\left(\mu, \sigma^2/n\right)$$

  (aproximadamente), onde $\mu \stackrel{\text{def}}{=} \mathbb{E}(X)$ e $\sigma^2 \stackrel{\text{def}}{=} \mathrm{Var}(X)$, desde que o tamanho da amostra, n, seja suficientemente grande.
- Agora, lembre que para qualquer variável aleatória Y, e dadas duas constantes a e b, temos que $\mathbb{E}(aY + b) = a\mathbb{E}(Y) + b$ e $\mathrm{Var}(aY + b) = a^2\,\mathrm{Var}(Y)$.
- Das duas propriedades acima discutidas, segue que

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

  (aproximadamente).

- Sabemos que, se $Z \sim N(0,1)$, então (ver Figura a seguir)
  (a) $\mathbb{P}(-1.645 \leq Z \leq 1.645) = 0.90$
  (b) $\mathbb{P}(-1.959 \leq Z \leq 1.959) = 0.95$
  (c) $\mathbb{P}(-2.576 \leq Z \leq 2.576) = 0.99$
- Logo temos, por exemplo,

$$\mathbb{P}\left(-1.645 \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq 1.645\right) \cong 0.90$$

e, portanto,

$$\mathbb{P}\left(\bar{X}_n - 1.645\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 1.645\frac{\sigma}{\sqrt{n}}\right) \cong 0.90$$

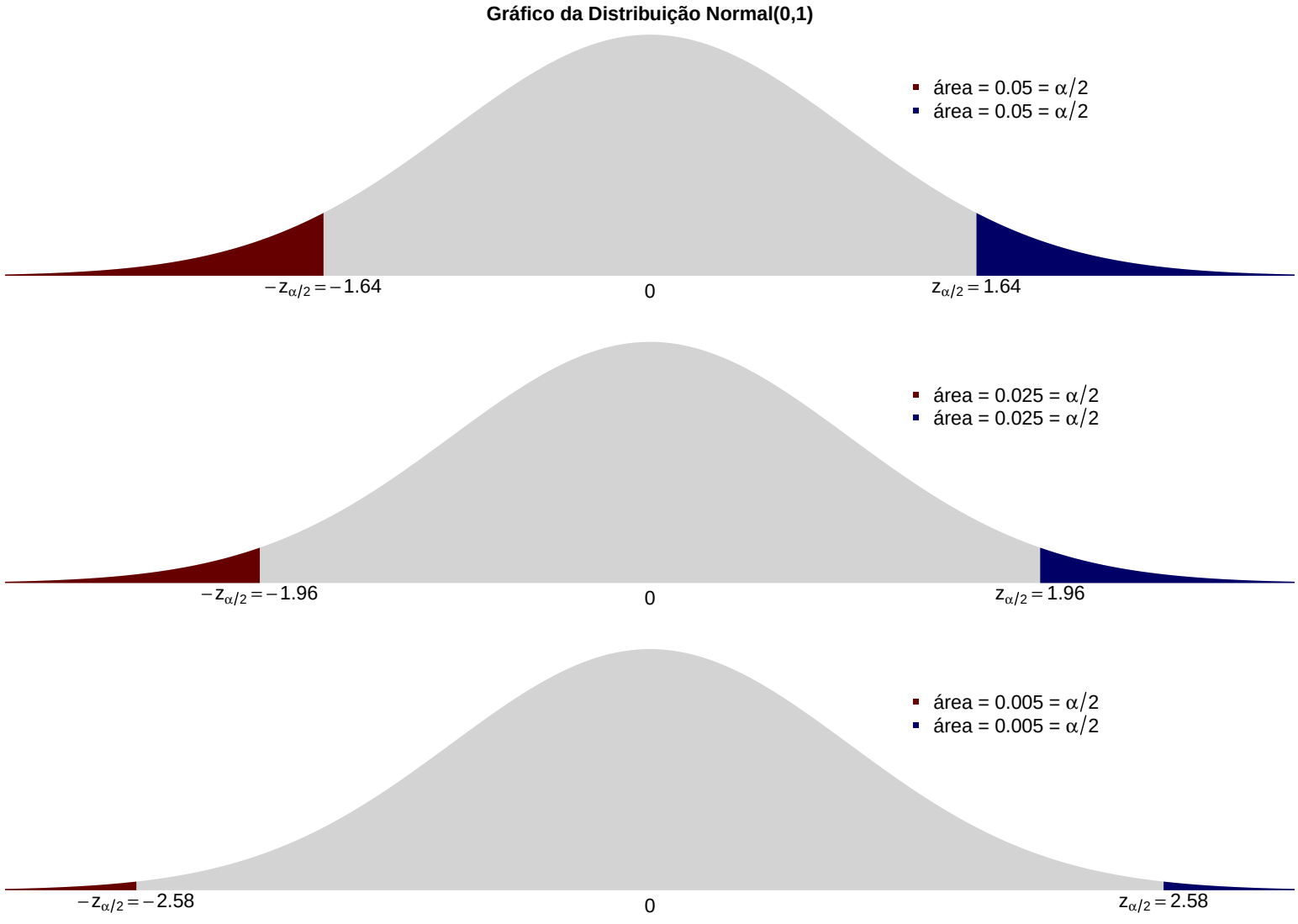


Gráfico da Distribuição Normal(0,1)
área = 0.05 = α/2
área = 0.05 = α/2
−z_α/2 = −1.64
0
z_α/2 = 1.64
área = 0.025 = α/2
área = 0.025 = α/2
−z_α/2 = −1.96
0
z_α/2 = 1.96
área = 0.005 = α/2
área = 0.005 = α/2
−z_α/2 = −2.58
0
z_α/2 = 2.58

- É necessário tomar cuidado com a interpretação das propriedades discutidas acima.
- O que mostramos no nosso exemplo é que é possível propor duas variáveis aleatórias

$$\ell = \bar{X}_n - 1.645\sigma/\sqrt{n} \qquad \text{e} \qquad L = \bar{X}_n + 1.645\sigma/\sqrt{n},$$

com $\ell < L$, tais que

$$\mathbb{P}([\ell \leq \mu] \cap [\mu \leq L]) = 0.90$$

- Agora suponha que o resultado do experimento tenha sido ω. Evidentemente, o intervalo $[\ell(\omega), L(\omega)]$ ou contém o verdadeiro parâmetro μ, ou não o contém.
- Dito de outra forma, $[\ell, L]$ é um intervalo aleatório, que em cada particular experimento ω pode conter ou não o parâmetro μ. O que podemos afirmar é que esse intervalo aleatório tem probabilidade igual a 0.90 de cobrir o verdadeiro parâmetro.

- Mais geralmente, vamos definir z_α pela equação

$$\mathbb{P}\big(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\big) = 1 - \alpha.$$

- Argumentando de maneira semelhante à do exemplo acima, concluímos que

$$\mathbb{P}\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \cong 1 - \alpha$$

- Se o resultado do experimento foi ω, dizemos que o intervalo

$$\left[\bar{X}_n(\omega) - z_{\alpha/2}\sigma/\sqrt{n},\, \bar{X}_n(\omega) + z_{\alpha/2}\sigma/\sqrt{n}\right]$$

é um **intervalo de confiança para o parâmetro μ, com coeficiente de confiança igual a $1 - \alpha$.**

- Notação:

$$\mathrm{IC}(\mu; 1 - \alpha) = \bar{X}_n(\omega) \pm z_{\alpha/2}\sigma/\sqrt{n}$$

- Em geral, podemos enquadrar os procedimentos de inferência em dois casos: 1) população normalmente distribuída; 2) demais casos.
- A ideia é caracterizar as distribuições amostrais das estatísticas

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{1}$$

e

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \tag{2}$$

em cada um desses casos.
- Conhecer a distribuição dessas estatísticas é importante para propormos intervalos de confiança e construirmos testes de hipóteses para o parâmetro μ.
- Se a variância populacional for conhecida (uma hipótese pouco realista mas útil do ponto de vista didático), utiliza-se a estatística (1). Caso contrário, utilizamos a estatística (2).

Caso 1 **A população é Normalmente distribuída**. Isto significa que a variável aleatória de interesse X segue uma distribuição Normal com uma certa média $\mathbb{E}(X) = \mu$ e variância $\mathrm{Var}(X) = \sigma^2$. Nesse caso, o estimador 'média amostral' tem distribuição *exatamente* Normal, com média μ e variância σ^2/n. Em outras palavras,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Ademais, vale que

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

quer dizer, a estatística padronizada segue uma distribuição student–t com $n-1$ graus de liberdade (essa é a distribuição exata dessa estatística). Obs.: se o tamanho da amostra for suficientemente grande, a distribuição t é muito parecida com a Normal(0,1).

Caso 2 **A população segue uma distribuição não–Normal**. Isto significa que a variável aleatória de interesse X segue alguma distribuição hipotética (que pode ou não ser especificada pelo pesquisador) com uma certa média $\mathbb{E}(X) = \mu$ e variância $\mathrm{Var}(X) = \sigma^2$. Nesse caso, o estimador 'média amostral' tem distribuição *aproximadamente* Normal, com média μ e variância σ^2/n. Em outras palavras,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{\text{aprox.}}{\sim} N(0,1),$$

desde que o tamanho da amostra, n, seja razoavelmente grande.

- Nesse contexto, contudo, a distribuição t não é util. O que temos é que a estatística (2) *também* é aproximadamente Normal:

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \overset{\text{aprox.}}{\sim} N(0,1),$$

desde que n seja suficientemente grande.

Testes de Hipóteses

Testes de Hipóteses

- Vamos introduzir as ideias e conceitos relacionados a Testes de Hipóteses a partir de um exemplo.
- Suponhamos que um navio vindo da Noruega chegou ao porto de Nova Iorque (estamos em meio à II Guerra Mundial e trata-se de um navio de refugiados!).
- Ninguém no navio fala inglês, e o fiscal de imigração não fala norueguês.
- Tudo que o fiscal sabe é que o navio é oriundo de uma das seguintes duas ilhas norueguesas: Tjøme ou Tromsøya. Todas as pessoas a bordo são habitantes da ilha de origem do navio.
- O fiscal dispõe de informações de um censo que contém dados sobre as alturas de toda a população norueguesa.
- A altura média (populacional) dos habitantes de Tjøme é igual a $\mu_A = 187$cm; a dos habitantes de Tromsøya é $\mu_B = 184$cm.
- O desvio padrão das alturas em Tjøme é de $\sigma_A = 10$cm. Em Tromsøya o desvio padrão é de $\sigma_B = 12$cm.

- Entre tripulação e passageiros, há 100 pessoas a bordo do navio.
- O fiscal tem ordens de seus superiores para acolher os refugiados caso estes sejam oriundos de Tromsøya, mas encaminhar o navio para New Haven – Connecticut, caso este seja oriundo de Tjøme.
- Em virtude de seu débil estado de saúde, John Fitzgerald (no caso, o fiscal) acumula 3 faltas no mês.
- Ao final do dia, seus superiores (que conhecem a procedência do navio) chegarão ao porto para a vistoria de rotina. Se o Sr Fitzgerald cometer o erro de acolher os refugiados de Tjøme, ele será demitido. Se redirecionar para New Haven o navio de Tromsøya, ele será suspenso por 1 mês, sem direito a salário.

- O fiscal de imigração quer testar a seguinte *hipótese*:

  Hipótese Nula (H_0): o navio é oriundo de Tjøme.

- O fiscal pode tão somente *rejeitar* essa hipótese, ou *não rejeitá-la*. Note que, se rejeitá-la, então o Sr Fitzgerald estará aceitando a

  Hipótese Alternativa (H_1): o navio é oriundo de Tromsøya.

- A escolha entre rejeitar ou não a hipótese nula será tomada a partir de uma *regra de decisão*, a qual será baseada em medições das alturas de tripulantes e passageiros do navio.
- Mais precisamente, o Sr Fitzgerald vai rejeitar sua hipótese (de que o navio é oriundo de Tjøme), se a altura média das 100 pessoas no navio for suficientemente baixa, pois isso lhe dará evidência de que na verdade essas 100 pessoas vieram da ilha de Tromsøya, onde a altura média da população é menor.

- Fundamentado em seu vasto conhecimento da ciência estatística, o Sr Fitzgerald elaborou a seguinte regra de decisão:

  Regra de decisão: “Rejeitar a hipótese nula se a altura média das 100 pessoas no navio for inferior a 185cm. Caso contrário, não rejeitar”.

- Se rejeitar a hipótese, o Sr Fitzgerald estará disposto a acreditar que o navio é oriundo de Tromsøya e, portanto, acolherá os refugiados.
- Por outro lado, se não rejeitar a hipótese, então ele acreditará que o navio é de fato oriundo de Tjøme, e redirecionará a embarcação para New Haven.
- Temos, portanto:

$$\begin{aligned} \text{Rejeitar} &\iff \text{Acolher} \\ \text{Não rejeitar} &\iff \text{Redirecionar} \end{aligned}$$

- Classificaremos os possíveis erros de decisão do Sr Fitzgerald da seguinte maneira:
  Erro do Tipo I: Rejeitar a hipótese nula, quando ela é verdadeira. Se o Sr Fitzgerald cometer esse erro, ele decidirá acreditar que o navio é oriundo de Tromsøya, quando na verdade é de Tjøme. Portanto, ele acolherá os refugiados quando deveria ter redirecionado-os para New Haven. Ao final do dia, nosso herói perderá o emprego.
  Erro do Tipo II: Não rejeitar a hipótese nula, quando a hipótese alternativa é verdadeira. Se o Sr Fitzgerald cometer esse erro, ele decidirá acreditar que o navio é oriundo de Tjøme, quando na verdade é de Tromsøya. Portanto, ele redirecionará os refugiados, quando deveria tê-los acolhido. Ao final do dia, nosso herói será suspenso por seus superiores.

Decisão / Origem	Tjøme	Tromsøya
Aceitar H_0	Decisão correta	Erro do Tipo II
Rejeitar H_0	Erro do Tipo I	Decisão correta

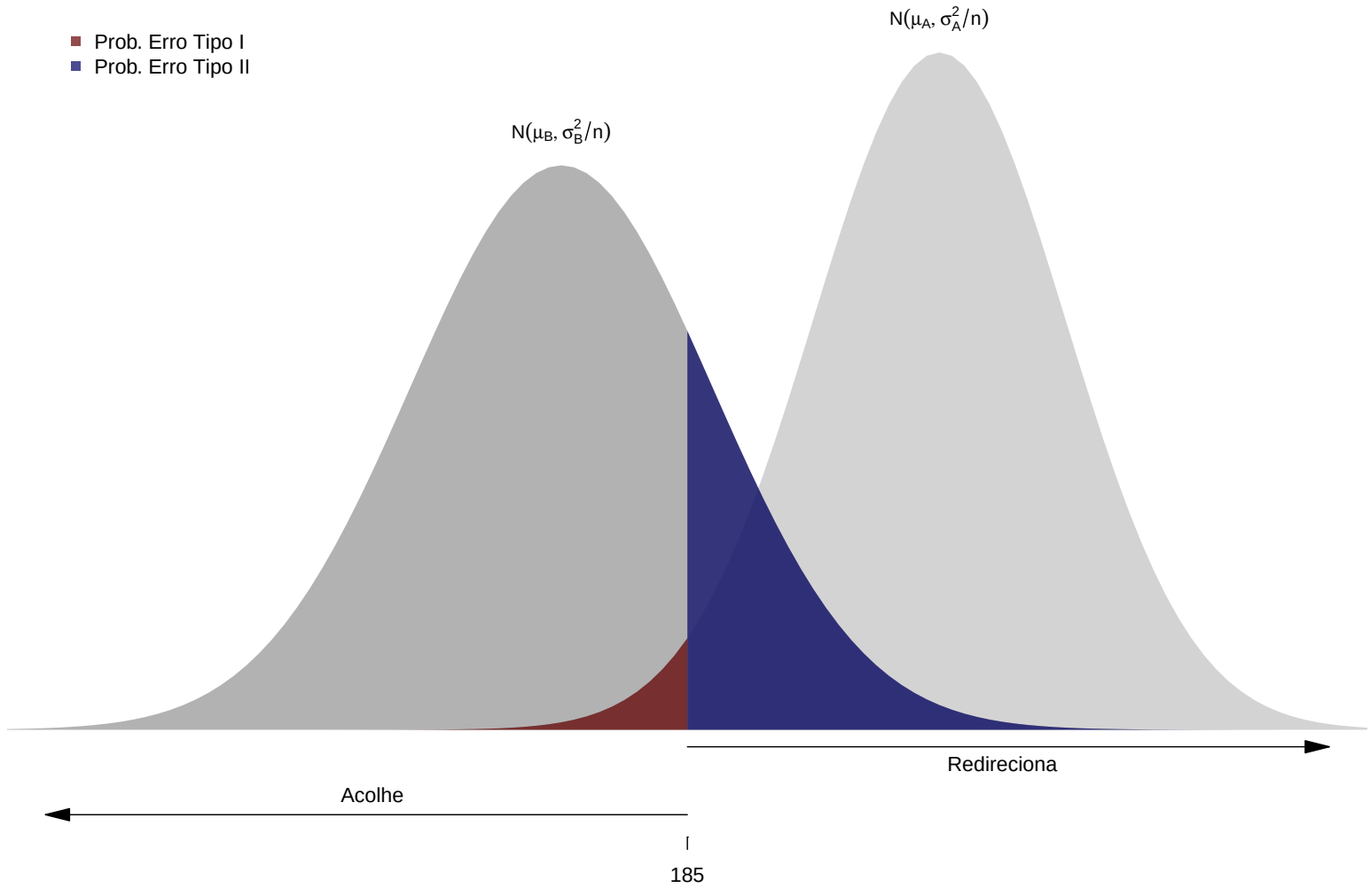
- Como o Sr Fitzgerald estabeleceu sua regra de decisão?
- Ele sabe que, se a embarcação é oriunda de Tjøme, então

$$\bar{X}_n \overset{\text{aprox.}}{\sim} N\left(\mu_A, \sigma_A^2/n\right).$$

- Por outro lado, se o navio partiu de Tromsøya, então

$$\bar{X}_n \overset{\text{aprox.}}{\sim} N\left(\mu_B, \sigma_B^2/n\right).$$

- Nesse exemplo, $n = 100$, e as aproximações distribucionais acima baseiam-se na suposição de que as pessoas que embarcaram no navio foram selecionadas ao acaso da população de sua ilha de origem.
- A Figura abaixo ilustra as duas possíveis distribuições populacionais referentes às hipóteses nula e alternativa.
- Fica claro que a *regra de decisão* do Sr Fitzgerald fica inteiramente determinada pela **Região crítica** $(-\infty, 185)$. A decisão de rejeitar ou não a hipótese nula depende unicamente de saber se a média (amostral) das alturas das 100 pessoas no navio situa-se ou não nessa região.


Prob. Erro Tipo I
Prob. Erro Tipo II
$N(\mu_A, \sigma_A^2/n)$
$N(\mu_B, \sigma_B^2/n)$
Redireciona
Acolhe
185

- E quais são as probabilidades de o Sr Fitzgerald cometer os Erros do Tipo I e II, respectivamente? Quanto ao Erro de Tipo I, temos

$$
\begin{aligned}
\mathbb{P}(\text{Erro do Tipo I}) &= \mathbb{P}(\text{Rejeitar } H_0 \text{ quando ela é verdadeira}) \\
&= \mathbb{P}(\text{Acolher os refugiados de Tjøme}) \\
&= \mathbb{P}(\text{Sr Fitzgerald ser demitido}) \\
&= \mathbb{P}(\bar{X}_n < 185 \,|\, H_0 \text{ é verdadeira}) \\
&= \mathbb{P}\left(\bar{X}_n < 185 \,|\, \bar{X}_n \overset{\text{aprox.}}{\sim} N\left(\mu_A, \sigma_A^2/n\right)\right) \\
&\cong 0.02275013 \cong 2,27\%
\end{aligned}
$$

- No que diz respeito ao Erro de Tipo II, por sua vez, temos

$$
\begin{aligned}
\mathbb{P}(\text{Erro do Tipo II}) &= \mathbb{P}(\text{Não rejeitar } H_0 \text{ quando } H_1 \text{ é verdadeira}) \\
&= \mathbb{P}(\text{Redirecionar o navio de Tromsøya}) \\
&= \mathbb{P}(\text{Sr Fitzgerald ser suspenso}) \\
&= \mathbb{P}(\bar{X}_n > 185 \,|\, H_1 \text{ é verdadeira}) \\
&= \mathbb{P}\left(\bar{X}_n > 185 \,|\, \bar{X}_n \overset{\text{aprox.}}{\sim} N\left(\mu_B, \sigma_B^2/n\right)\right) \\
&\cong 0.2023284 \cong 20.23\%
\end{aligned}
$$

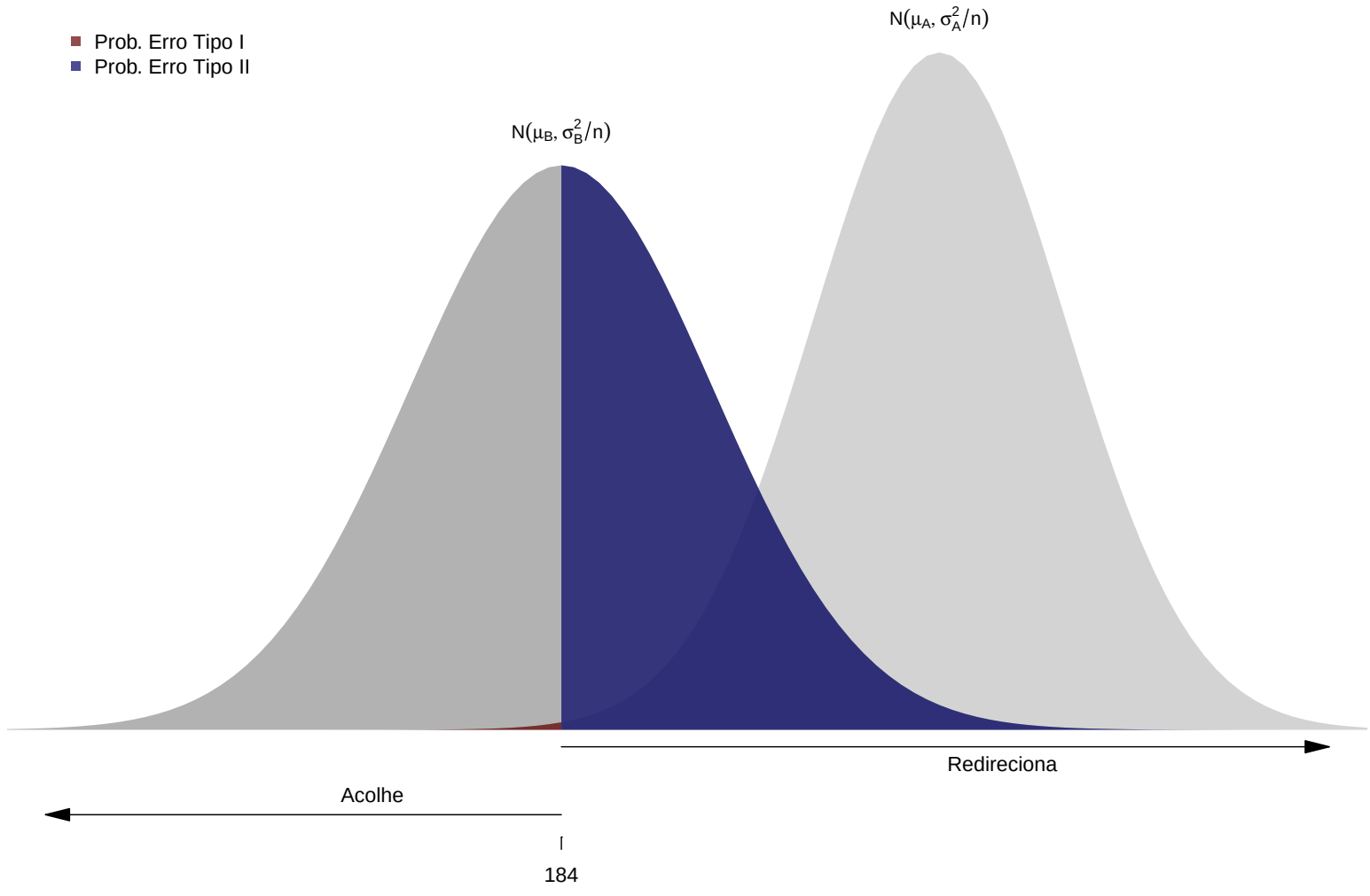
- O quê aconteceria se o Sr Fitzgerald quisesse reduzir a chance de ser demitido?
- Digamos que ele adote uma outra regra de decisão:

  Regra de decisão': "Rejeitar a hipótese nula se a altura média das 100 pessoas no navio for inferior a 184cm. Caso contrário, não rejeitar".

- Nesse caso, teríamos

$$\mathbb{P}(\text{Erro do Tipo I}) \cong 0.13\% \qquad \mathbb{P}(\text{Erro do Tipo II}) \cong 50\%$$

- Assim, o Sr Fitzgerald estaria praticamente certo de que não será demitido, mas as chances de que ele seja suspenso aumentariam significativamente! Veja a Figura abaixo.


Prob. Erro Tipo I
Prob. Erro Tipo II
N(μA, σ2A/n)
N(μB, σ2B/n)
Redireciona
Acolhe
184

Procedimento Geral do Teste de Hipóteses

- Podemos abstrair as ideias gerais de um teste de hipóteses sobre um parâmetro populacional da seguinte forma.
- Considere uma certa variável X associada a dada população.
- Suponha que tenhamos uma hipótese sobre um parâmetro populacional θ. Nesse contexto, uma **hipótese** é sempre uma *afirmação* sobre o parâmetro. Por exemplo: "a altura média populacional dos habitantes de Tromsøya é igual a 182cm".
- Mais geralmente, poderíamos ter uma afirmação do tipo: "o valor do parâmetro θ em questão é igual a θ_0".
- Tal afirmação é necessariamente ou verdadeira, ou falsa.
- (Note que no segundo exemplo acima, θ é o 'nome' do parâmetro, enquanto θ_0 é o valor numérico que estamos conjecturando)

- Iniciamos a análise explicitando qual é a nossa hipótese inicial, chamada **hipótese nula**:

$$H_0 : \theta = \theta_0$$

- Coletamos uma amostra $X_1, \ldots, X_n$ da variável X. Utilizando a evidência trazida por essa amostra, temos duas possibilidades:
  1. Rejeitar H_0
  2. Não rejeitar H_0
- É conveniente explicitar qual será a hipótese que vamos considerar aceitável, caso rejeitemos a hipótese nula. A essa hipótese chamamos de **hipótese alternativa**.
- No caso mais geral, a hipótese alternativa seria

$$H_1 : \theta \neq \theta_0$$

- Todavia, podemos considerar hipóteses alternativas um pouco mais restritivas. Por exemplo,

$$H_1 : \theta > \theta_0$$

  e da mesma forma com $\theta < \theta_0$ ao invés de $\theta > \theta_0$.
- A nossa decisão de rejeitar ou não H_0 será tomada de acordo com uma regra de decisão: uma **regra de decisão** é um procedimento binário que consiste em observar se uma **estatística de teste** (isto é, um estimador) situa-se ou não em uma **região crítica** (também chamada de **região de rejeição**).
- Se $\hat{\theta}$ é uma estatística de teste e RC é uma região crítica pré–especificada, então a regra de decisão fica inteiramente determinada pelo seguinte procedimento:

$$\begin{cases} \text{rejeitar } H_0 \text{ se } \hat{\theta}(\omega) \in RC \\ \text{não rejeitar } H_0 \text{ se } \hat{\theta}(\omega) \notin RC \end{cases}$$

- Um aspecto crucial é que, sob a validade da hipótese nula, a distribuição de probabilidade da estatística $\hat{\theta}$ seja conhecida.
- Todavia, nem sempre conhecemos a distribuição da estatística de teste *sob a hipótese alternativa*.
- No nosso exemplo anterior, a estatística de teste $\bar{X}_n$ tinha sua distribuição conhecida tanto sob a hipótese nula quanto sob a alternativa.

- Se alterássemos um pouco o nosso exemplo, adicionando outras possíveis ilhas de procedência para o navio (digamos, ilhas C e D), e adotássemos as seguintes hipóteses nula e alternativa, respectivamente:

$$H_0: \text{o navio é oriundo de Tjøme}$$
$$H_1: \text{o navio não é oriundo de Tjøme}$$

  então saberíamos que sob H_0 a estatística de teste $\bar{X}_n$ tem distribuição aproximadamente Normal com média μ_A e variância σ_A^2/n. Contudo, sob a H_1, a distribuição de $\bar{X}_n$ não está bem especificada (podemos até conhecer sua distribuição para cada possível ilha, mas isso não está explicitado por H_1).

- Portanto, no caso geral sempre podemos calcular a seguinte probabilidade

$$\mathbb{P}\Big(\hat{\theta} \in RC \,|\, H_0 \text{ é verdadeira}\Big) = \alpha.$$

- Essa é a probabilidade de se cometer o **Erro do Tipo I**: rejeitar a hipótese nula, quando ela é verdadeira.
- Todavia, quanto ao **Erro do Tipo II** (não rejeitar H_0 quando ela é falsa), nem sempre será possível calcular sua probabilidade.
- A probabilidade α é chamada o **nível de significância do teste**.
- Em geral, o que fixamos a priori é o nível de significância, não a região crítica. Esta é usualmente obtida *após* fixado o nível do teste, de forma que a identidade acima se verifique (em outras palavras, a região crítica depende de α, quer dizer, $RC = RC(\alpha)$.

Teste sobre a média de uma população com variância conhecida

- Suponhamos que uma certa variável X associada a uma população, seja tal que $\mathbb{E}(X) = \mu$ (desconhecido) e $\mathrm{Var}(X) = \sigma^2$ (conhecido).
- A partir de uma amostra $X_1, \ldots, X_n$ dessa variável, desejamos testar a hipótese nula de que a média populacional μ é igual a um certo número pré–fixado μ_0.
- A decisão de rejeitar ou não a hipótese nula será tomada considerando-se o valor da estatística de teste $\bar{X}_n$.
- Vamos fixar um valor de α, onde $0 < \alpha < 1$. Queremos obter uma região crítica RC tal que α seja precisamente a probabilidade de cometermos o Erro do Tipo I.

- Se H_0 for verdadeira, então vale (ao menos aproximadamente)

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

- Seja Z uma variável aleatória $N(0,1)$. Tomando z_α tal que $\mathbb{P}(Z > z_\alpha) = \alpha$ (e portanto também vale que $\mathbb{P}(Z < -z_\alpha) = \alpha$, concluímos que a região crítica

$$RC = \left\{x\colon\ x \leq \mu_0 - \tfrac{\sigma}{\sqrt{n}} z_{\alpha/2} \text{ ou } x \geq \mu_0 + \tfrac{\sigma}{\sqrt{n}} z_{\alpha/2}\right\}$$

  tem a propriedade desejada, isto é,

$$\mathbb{P}(\bar{X}_n \in RC \,|\, H_0 \text{ é verdadeira}) = \alpha.$$

- A nossa regra de decisão é, consequentemente, a de rejeitar a hipótese nula se $\bar{X}_n(\omega) \leq \mu_0 - \sigma z_{\alpha/2}/\sqrt{n}$ ou se $\bar{X}_n(\omega) \geq \mu_0 + \sigma z_{\alpha/2}/\sqrt{n}$.
- Com essa regra de decisão, estaremos cometendo o Erro do Tipo I com probabilidade igual a α.

- **Importante**: note que a construção do teste de hipóteses não depende da amostra!
- Quer dizer, antes de coletar os dados / amostra, já temos definidas as hipóteses nula e alternativa, a região crítica e a regra de decisão.
- Os dados só serão utilizados na última etapa do procedimento, onde simplesmente observaremos se a estatística de teste se situa ou não na região crítica pré-estabelecida, assim rejeitando ou não a hipótese nula.

- **Exemplo**: Suponhamos que uma máquina de preencher pacotes de café o faça de acordo com uma distribuição Normal com média μ e variância $\sigma^2 = 400\text{mg}^2$.
- Quer dizer, se X é a variável aleatória "peso de um pacote preenchido pela máquina", então $X \sim N(\mu, 20^2)$.
- Diremos que a a produção está sob controle se $\mu = 500\text{mg}$.
- Será coletada uma amostra de $n = 16$ pacotes preenchidos por essa máquina, e a estatística de teste 'média amostral' correspondente será utilizada para rejeitar ou não a hipótese nula de que a produção está sob controle.
- As hipóteses nula e alternativa são:

$$H_0 : \mu = 500$$
$$H_1 : \mu \neq 500.$$

- Este é um exemplo de teste bilateral, e nesse caso é o mais apropriado visto que estaremos interessados em rejeitar H_0 tanto se a a estatística de teste for muito alta quanto se for muito baixa (indicativos de que a produção não está sob controle).

- Vamos fixar o nível de significância do teste $\alpha = 0.01 = 1\%$.
- Agora podemos obter a região crítica do teste: temos, sob H_0, que

$$\bar{X}_n \sim N\big(500,\, 20^2/16\big)$$

- Dito de outra forma, sob H_0 vale que

$$\frac{\bar{X}_n - 500}{20/4} \sim N(0,1)$$

- Assim, lembrando que z_a é o número que soluciona a equação $\mathbb{P}(Z > z_a) = a$, onde $Z \sim N(0,1)$, temos para $a = \alpha/2 = 0.005$ que $z_{\alpha/2} = 2.58$.
- Quer dizer,

$$\mathbb{P}\big(\bar{X}_n < 500 - 2.58\tfrac{20}{4} \text{ ou } \bar{X}_n > 500 + 2.58\tfrac{20}{4}\big) = 0.01$$

- Portanto, a região crítica que adotaremos será

$$RC = (-\infty, 478.1) \cup (512.9, +\infty).$$

- Em outras palavras, rejeitaremos a hipótese nula se a estatística de teste tiver um valor, na amostra que coletaremos, inferior a 478.1 ou superior a 512.9
- Digamos que, em uma particular amostra ω obtivemos o valor $\bar{X}_{16}(\omega) = 492$.
- Nesse caso, não vamos rejeitar a hipótese nula: não há evidência suficiente nos dados (representada pelo valor da estatística de teste) para rejeitarmos a nossa hipótese de que a média populacional é igual a 500mg, isto é, de que a produção está sob controle.

Teste sobre a média de uma população com variância desconhecida

- Quando a variância populacional é desconhecida, utilizamos a estatística de teste

$$\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

  onde $S_n^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$, isto é, substituímos o valor de σ por um estimador, S_n.
- Se a amostra $X_1, \ldots, X_n$ é iid de uma população $N(\mu, \sigma^2)$, então sob $H_0 : \mu = \mu_0$ a estatística de teste acima tem distribuição t de Student, com $n-1$ graus de liberdade.
- A região crítica para testar H_0 é obtida, portanto, a partir dos valores críticos da distribuição $t(n-1)$.